A Gentle Introduction to Dimension Reduction and Semiparametric Approach

Yanyuan Ma

University of South Carolina

December 11, 2015

Outline

Two Assumptions to Reduce Dimension

- One aspect of data complexity: number of covariates
- Assumption 1: Sparsity
  - Only a few of the covariates useful
  - Variable selection, often via penalization
- Assumption 2: Reducibility
  - Only a few linear combinations useful
  - (Sufficient) Dimension reduction

Specific Dimension Reduction Models

- Notation: $Y \in R$, $\mathbf{x} \in R^{p \times 1}$, $\beta \in R^{p \times d}, p > d$.
- Central space model
- Central mean space model
- Central moment space model
- Central variance space model
- Central median space model
- Central quantile space model

Central Space Model

- Central space model assumption:
  Distribution of Y relates to covariates $\mathbf{x}$ only through $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}$.
- Mathematical form:

$$Y \perp\!\!\!\perp \mathbf{x} \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}$$

  or equivalently

$$\mathrm{pr}(Y \leq y \mid \mathbf{x}) = \mathrm{pr}(Y \leq y \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}) \text{ for all } y.$$

- Central space $\mathcal{S}_{Y|\mathbf{x}}$: Span of the columns in $\boldsymbol{\beta}$.
- Goal: Estimate $\mathcal{S}_{Y|\mathbf{x}}$

Central Mean Space Model

- Central mean space model assumption:
  Mean of Y relates to covariates $\mathbf{x}$ only through $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}$.
- Mathematical form:

$$E(Y \mid \mathbf{x}) = E(Y \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})$$

  or equivalently

$$Y = m(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}) + \epsilon, \quad E(\epsilon \mid \mathbf{x}) = 0.$$

- Central mean space $\mathcal{S}_{E(Y|\mathbf{x})}$: Span of the columns in $\boldsymbol{\beta}$.
- Goal: Estimate $\mathcal{S}_{E(Y|\mathbf{x})}$

Central Moment Space Model

- Central moment space model assumption:
  First k moments of Y relates to covariates $\mathbf{x}$ only through $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}$.
- Mathematical form:

$$E(Y^j \mid \mathbf{x}) = E(Y^j \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}) \text{ for } j = 1, \ldots, k.$$

  or equivalently

$$Y^j = m_j(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}) + \epsilon_j, \quad E(\epsilon_j \mid \mathbf{x}) = 0 \text{ for } j = 1, \ldots, k.$$

- Central k-moment space $\mathcal{S}^{(k)}_{Y|\mathbf{x}}$: Span of the columns in $\boldsymbol{\beta}$.
- Goal: Estimate $\mathcal{S}^{(k)}_{Y|\mathbf{x}}$

Central Variance Space Model

- Central variance space model assumption:
  Conditional variance of Y on $\mathbf{x}$ is a function of $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}$ only.
- Mathematical form:

$$\mathrm{var}(Y \mid \mathbf{x}) = E\{\mathrm{var}(Y \mid \mathbf{x}) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}\}$$

- It is not equivalent to

$$\mathrm{var}(Y \mid \mathbf{x}) = \mathrm{var}(Y \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})$$

- Central variance space $\mathcal{S}_{\mathrm{var}(Y|\mathbf{x})}$: Span of the columns in $\boldsymbol{\beta}$.
- Goal: Estimate $\mathcal{S}_{\mathrm{var}(Y|\mathbf{x})}$

Estimation in Dimension Reduction Models

- Target: Finding the column space of β with the smallest d.
  - Estimating space, not parameters
  - Smallest space exists and unique
- Invariance: Let $\mathbf{z} = \mathbf{A}^{-1}(\mathbf{x} - \mu)$, then

$$\begin{aligned}
\mathrm{pr}(Y \mid \beta^{\mathrm{T}}\mathbf{x}) &= \mathrm{pr}\{Y \mid (\mathbf{A}\beta)^{\mathrm{T}}\mathbf{z}\}, \\
\mathrm{pr}(Y^j \mid \beta^{\mathrm{T}}\mathbf{x}) &= \mathrm{pr}\{Y^j \mid (\mathbf{A}\beta)^{\mathrm{T}}\mathbf{z}\}, \\
\mathrm{var}(Y \mid \beta^{\mathrm{T}}\mathbf{x}) &= \mathrm{var}\{Y \mid (\mathbf{A}\beta)^{\mathrm{T}}\mathbf{z}\},
\end{aligned}$$

  which leads to

$$\mathcal{S}_{Y|\mathbf{x}} = \mathbf{A}^{-1}\mathcal{S}_{Y|\mathbf{z}},\ \mathcal{S}_{E(Y|\mathbf{x})} = \mathbf{A}^{-1}\mathcal{S}_{E(Y|\mathbf{z})},$$
$$\mathcal{S}^{(k)}_{Y|\mathbf{x}} = \mathbf{A}^{-1}\mathcal{S}^{(k)}_{Y|\mathbf{z}},\ \mathcal{S}_{\mathrm{var}(Y|\mathbf{x})} = \mathbf{A}^{-1}\mathcal{S}_{\mathrm{var}(Y|\mathbf{z})}$$

- Simplification: $E(\mathbf{x}) = \mathbf{0}$, $\mathrm{cov}(\mathbf{x}) = \mathbf{I}_p$.

Three Classes of Estimation Approaches

1. Inverse regression based methods.
   - $\mathcal{S}_{Y|\mathbf{x}}$: SIR, SAVE, DR
   - $\mathcal{S}_{E(Y|\mathbf{x})}$: OLS, PHD
2. Nonparametric estimation:
   - $\mathcal{S}_{Y|\mathbf{x}}$: dMAVE
   - $\mathcal{S}_{E(Y|\mathbf{x})}$: MAVE
3. Semiparametric estimation:
   - $\mathcal{S}_{Y|\mathbf{x}}$: Semi-SIR, Semi-DR, Semi-SAVE, Efficient
   - $\mathcal{S}_{E(Y|\mathbf{x})}$: Semi-PHD, Local Efficient, Efficient

Inverse Regression Based Methods

- Central idea: Invert the position between Y and $\mathbf{x}$
- Requirements:
  - Linearity condition

$$E\left(\mathbf{x} \mid \beta^{\mathrm{T}}\mathbf{x}\right) = \mathbf{P}\mathbf{x} = \beta(\beta^{\mathrm{T}}\beta)^{-1}\beta^{\mathrm{T}}\mathbf{x}$$

  - Constant variance condition

$$\mathrm{cov}\left(\mathbf{x} \mid \beta^{\mathrm{T}}\mathbf{x}\right) = \mathbf{Q} = \mathbf{I}_p - \mathbf{P}$$

- Linearity/constant variance condition need to hold at true β
- Linearity condition approximately holds for large p, fixed d.
- Hard to check. If strengthened to any β: elliptical/normal distribution

Sliced Inverse Regression (SIR)

- SIR estimate of $\mathcal{S}_{Y|\mathbf{x}}$: non-zero eigen-space of $\mathbf{\Lambda}_{\text{sir}} = \text{cov}\{E(\mathbf{x} \mid Y)\}$.
- Procedure:
  - Estimate $\widehat{E}(\mathbf{x} \mid Y)$: p univariate problems, nonparametric regression or average in slices
  - Form sample variance-covariance of $\widehat{E}(\mathbf{x} \mid Y)$.
  - Extremely simple procedure!
- SIR requirement: critically relies on linearity condition.

Sliced Average Variance Estimation (SAVE)

- SAVE estimate of $\mathcal{S}_{Y|\mathbf{x}}$: non-zero eigen-space of $\mathbf{\Lambda}_{\text{save}} = E\left[\{\mathbf{I}_p - \text{cov}(\mathbf{x} \mid Y)\}^2\right]$.
- Procedure:
  - Estimate $\widehat{\text{cov}}(\mathbf{x} \mid Y)$: $O(p^2)$ univariate problems, nonparametric regression or sample variance-covariance in slices
  - Form sample average of $\left\{\mathbf{I}_p - \widehat{\text{cov}}(\mathbf{x} \mid Y)\right\}^2$.
  - Computationally simple
- SAVE requirement: critically relies on linearity condition and constant variance condition.

SIR or SAVE?

- SAVE more comprehensive
- SIR more efficient
- Convex combination of SIR and SAVE

$$\alpha \mathbf{\Lambda}_{\text{sir}} + (1 - \alpha)\mathbf{\Lambda}_{\text{save}}, \quad 0 \leq \alpha \leq 1$$

- More comprehensive than SIR, more efficient than SAVE.

Directional Regression (DR)

- DR estimate of $\mathcal{S}_{Y|\mathbf{x}}$: non-zero eigen-space of
  $\Lambda_{\mathrm{dr}} = E[\{2\mathbf{I}_p - \mathbf{A}(Y, \widetilde{Y})\}^2]$,
  $\mathbf{A}(Y, \widetilde{Y}) = E\{(\mathbf{x} - \widetilde{\mathbf{x}})(\mathbf{x} - \widetilde{\mathbf{x}})^{\mathrm{T}} \mid Y, \widetilde{Y}\}$
  $(\widetilde{\mathbf{x}}, \widetilde{Y})$: an independent copy of $(\mathbf{x}, Y)$.
- Procedure:
  - Estimate $\widehat{E}\{(\mathbf{x} - \widetilde{\mathbf{x}})(\mathbf{x} - \widetilde{\mathbf{x}})^{\mathrm{T}} \mid Y, \widetilde{Y}\}$: $O(p^2)$ univariate problems, nonparametric regression or sample average in slices
  - Form sample average of $\{2\mathbf{I}_p - \widehat{\mathbf{A}}(Y, \widetilde{Y})\}^2$.
  - Improve both SIR and SAVE
- DR requirement: critically relies on linearity condition and constant variance condition.

Ordinary Least Squares (OLS)

- OLS estimate of $\mathcal{S}_{E(Y|\mathbf{x})}$:
  - Perform ordinary least square on model $Y = \beta^{\mathrm{T}}\mathbf{X} + \epsilon$
  - The column space of $\widehat{\beta}$ belongs to $\mathcal{S}_{E(Y|\mathbf{x})}$.
- Finds one direction, not all directions of $\mathcal{S}_{E(Y|\mathbf{x})}$.
- OLS requirement: critically relies on linearity condition

Principal Hessian Direction (PHD)

- PHD estimate of $\mathcal{S}_{E(Y|\mathbf{x})}$: non-zero eigen-space of $\mathbf{\Lambda}_{\text{phd}} = E\left[\{Y - E(Y)\}\,\mathbf{x}\mathbf{x}^{\mathrm{T}}\right]$
- Procedure:
  - Center Y
  - Form sample average of $\left\{Y - \widehat{E}(Y)\right\}\mathbf{x}\mathbf{x}^{\mathrm{T}}$.
  - Extremely simple computation, finds all directions in $\mathcal{S}_{E(Y|\mathbf{x})}$.
- PHD requirement: critically relies on linearity condition and constant variance condition.

Notes on Inverse Regression Estimators

- Elegant and mysterious
- Issues in tuning parameter selection: number of slices or bandwidth
- More and extended inverse regression estimators available both for $\mathcal{S}_{Y|\mathbf{x}}$ and $\mathcal{S}_{E(Y|\mathbf{x})}$
- Linearity/constant variance conditions bring some constraints
- Transformation often applied to reach these conditions
- Further assuming parametric model on $\mathbf{x} \mid Y$ bypasses high dimensional issue, but assumption too strong

Nonparametric Methods

- General idea of nonparametric methods: Minimize a criterion that describes the goodness-of-fit of the model.
- MAVE and dMAVE
  - Least square criterion
  - Nonparametric estimation involved: Local linear estimator
- Requirement: All components in $\mathbf{x}$ are continuous

Minimum Average Variance Estimation (MAVE)

- Local linear estimator for $m(\mathbf{z})$ in $Y = m(\mathbf{z}) + \epsilon$

$$\min_{a,\mathbf{b}} \sum_{i=1}^{n} \{Y_i - a - \mathbf{b}^{\mathrm{T}}(\mathbf{z}_i - \mathbf{z})\}^2 K_h(\mathbf{z}_i - \mathbf{z})$$

- $\mathcal{S}_{E(Y|\mathbf{x})}$ model $Y = m(\beta^{\mathrm{T}}\mathbf{x}) + \epsilon$
- MAVE for $\mathcal{S}_{E(Y|\mathbf{x})}$
  - let $w_{ij} = K_h(\beta^{\mathrm{T}}\mathbf{x}_i - \beta^{\mathrm{T}}\mathbf{x}_j)$.
  - Set $\mathbf{z} = \beta^{\mathrm{T}}\mathbf{x}_j$'s.

$$\min \sum_{j=1}^{n} \sum_{i=1}^{n} \{Y_i - a_j - \mathbf{b}_j^{\mathrm{T}}(\beta^{\mathrm{T}}\mathbf{x}_i - \beta^{\mathrm{T}}\mathbf{x}_j)\}^2 w_{ij}$$

- Other criterion possible
- Other nonparametric estimator possible

Density Based MAVE (dMAVE)

- dMAVE for $\mathcal{S}_{Y|\mathbf{x}} : f_{Y|\mathbf{x}}(y, \mathbf{x}) = f_{Y|\beta^{\mathrm{T}}\mathbf{x}}(y, \beta^{\mathrm{T}}\mathbf{x})$
  - View it as $I(Y = y) = m(\beta^{\mathrm{T}}\mathbf{x}) + \epsilon$ for all y values.
  - MAVE yields

$$\min \sum_{j=1}^{n} \sum_{i=1}^{n} \{K_b(Y_i - y) - a_j - \mathbf{b}_j^{\mathrm{T}}(\beta^{\mathrm{T}}\mathbf{x}_i - \beta^{\mathrm{T}}\mathbf{x}_j)\}^2 w_{ij}$$

  - Select $y = Y_1, \ldots, Y_n$

$$\min \sum_{k=1}^{n} \sum_{j=1}^{n} \sum_{i=1}^{n} \{K_b(Y_i - Y_k) - a_{jk} - \mathbf{b}_{jk}^{\mathrm{T}}(\beta^{\mathrm{T}}\mathbf{x}_i - \beta^{\mathrm{T}}\mathbf{x}_j)\}^2 w_{ij}$$

- Other criterion possible
- Other nonparametric estimator possible

Notes on Nonparametric Estimators

- Intuitive and computationally intensive
- Tuning parameter selected via crossvalidation
- More and extended nonparametric estimators available both for $\mathcal{S}_{Y|\mathbf{x}}$ and $\mathcal{S}_{E(Y|\mathbf{x})}$
- Do not need linearity/constant variance conditions
- Need continuity of each covariate. Could be relaxed with more computational complexity
- Often yields better performance than inverse regression based methods

Semiparametric Methods

- $\mathcal{S}_{Y|\mathbf{x}}$ model is a semiparametric model:

$$f_{X,Y}(\mathbf{x}, Y; \boldsymbol{\beta}, \eta) = \eta_1(\mathbf{x})\eta_2(Y, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})$$

  $\boldsymbol{\beta}$: of interest; η_1, η_2: nuisance
- Regular Asymptotically Linear (RAL) estimator:

$$n^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = n^{-1/2}\sum_{i=1}^{n}\phi(\mathbf{O}_i) + o_p(1)$$

  $\phi(\mathbf{O}_i)$: the i-th influence function.
- Estimator: $\sum_{i=1}^{n}\phi(\mathbf{O}_i, \widehat{\boldsymbol{\beta}}) = 0, \mathrm{var}(\sqrt{n}\widehat{\boldsymbol{\beta}}) = \mathrm{E}\{\phi(\mathbf{O})\phi(\mathbf{O})^{\mathrm{T}}\}$
- Find influence function $\Longrightarrow$ find RAL estimator

The Geometry

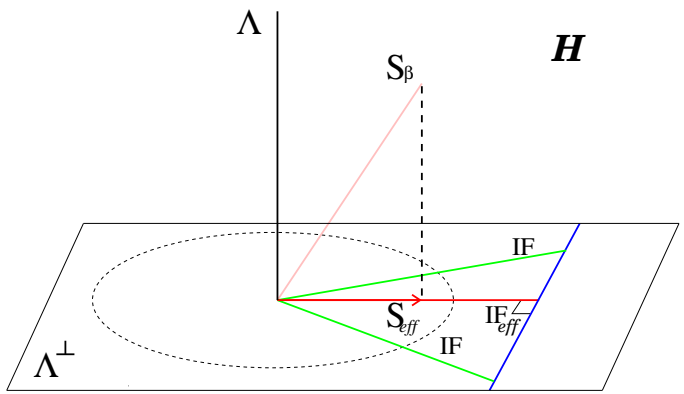
- Hilbert space

$$\mathcal{H}(\mathbf{O}) = [\mathbf{f}(\mathbf{O}) : \mathrm{E}\{\mathbf{f}(\mathbf{O})\} = 0, \mathrm{var}\{\mathbf{f}(\mathbf{O})\} \text{ finite}]$$
$$< \mathbf{f}(\mathbf{O}), \mathbf{g}(\mathbf{O}) >= \mathrm{cov}\{\mathbf{f}(\mathbf{O}), \mathbf{g}(\mathbf{O})\} = E\{\mathbf{f}(\mathbf{O})^{\mathrm{T}}\mathbf{g}(\mathbf{O})\}$$

- Λ for parametric model: $p(\mathbf{O}; \beta, \gamma)$
  - Nuisance score function: $\mathbf{S}_\gamma(\mathbf{O}) = \partial \log p(\mathbf{O}, \beta, \gamma)/\partial\gamma$
  - Nuisance tangent space $\Lambda_\gamma \subset \mathcal{H}$: a linear space spanned by nuisance score function $\mathbf{S}_\gamma$
- Λ for semiparametric model: $p(\mathbf{O}; \beta, \eta)$
  - Parametric submodel $p(\mathbf{O}; \beta, \gamma)$
  - Nuisance tangent space Λ: span of $\mathbf{S}_\gamma$ for all γ's
- $\Lambda^\perp$: Orthogonal complement of Λ

The Influence Functions

- Two requirements for influence function:
  1. $\phi(\mathbf{O}) \in \mathbf{\Lambda}^{\perp}$
  2. $E\{\phi(\mathbf{O})\mathbf{S}_{\boldsymbol{\beta}}^{\mathrm{T}}(\mathbf{O})\} = \mathbf{I}$,
     $\mathbf{S}_{\boldsymbol{\beta}}(\mathbf{O}) = \partial \log p(\mathbf{O})/\partial \boldsymbol{\beta}$ is the score vector
- The spaces and influence functions visually



- Efficient influence function:

$$\phi_{\mathrm{eff}}(\mathbf{O}) = \mathrm{E}\{\mathbf{S}_{\mathrm{eff}}(\mathbf{O})\mathbf{S}_{\mathrm{eff}}(\mathbf{O})^{\mathrm{T}}\}^{-1}\mathbf{S}_{\mathrm{eff}}(\mathbf{O})$$

$\Lambda^{\perp}$ for $\mathcal{S}_{Y|\mathbf{x}}$ and Many Possibilities

- $\Lambda = \Lambda_1 \oplus \Lambda_2$, where

$$\begin{aligned}\Lambda_1 &= \{\mathbf{f}(\mathbf{x}) : \forall \mathbf{f} \text{ such that } E(\mathbf{f}) = \mathbf{0}\} \\ \Lambda_2 &= \{\mathbf{f}(Y, \beta^{\mathrm{T}}\mathbf{x}) : \forall \mathbf{f} \text{ such that } E(\mathbf{f} \mid \mathbf{x}) = E(\mathbf{f} \mid \beta^{\mathrm{T}}\mathbf{x}) = \mathbf{0}\}.\end{aligned}$$

- $\Lambda^{\perp} = \{\mathbf{f}(Y, \mathbf{x}) - E(\mathbf{f} \mid \beta^{\mathrm{T}}\mathbf{x}, Y) : E(\mathbf{f} \mid \mathbf{x}) = E(\mathbf{f} \mid \beta^{\mathrm{T}}\mathbf{x}) \forall \mathbf{f}\}$.
- $\Lambda^{\perp}$ provides many candidate influence functions (unnormalized)
  - For any $\mathbf{g}, \mathbf{a}$,

$$\left\{\mathbf{g}\left(Y, \beta^{\mathrm{T}}\mathbf{x}\right) - E\left(\mathbf{g} \mid \beta^{\mathrm{T}}\mathbf{x}\right)\right\}\left\{\alpha\left(\mathbf{x}\right) - E\left(\alpha \mid \beta^{\mathrm{T}}\mathbf{x}\right)\right\}$$

  - For any $\mathbf{g}_i, \mathbf{a}_i$,

$$\sum_{i=1}^{k}\left\{\mathbf{g}_i\left(Y, \beta^{\mathrm{T}}\mathbf{x}\right) - E\left(\mathbf{g}_i \mid \beta^{\mathrm{T}}\mathbf{x}\right)\right\}\left\{\alpha_i\left(\mathbf{x}\right) - E\left(\alpha_i \mid \beta^{\mathrm{T}}\mathbf{x}\right)\right\}$$

Double Centering Form and Double Robustness

- Original form

$$\left\{\mathbf{g}\left(Y, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}\right) - E\left(\mathbf{g} \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}\right)\right\}\left\{\boldsymbol{\alpha}\left(\mathbf{x}\right) - E\left(\boldsymbol{\alpha} \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}\right)\right\}$$

- Mis-specify $E\left(\boldsymbol{\alpha} \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}\right)$ does not cause inconsistency.

$$E\left[\left\{\mathbf{g}\left(Y, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}\right) - E\left(\mathbf{g} \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}\right)\right\}\left\{\boldsymbol{\alpha}\left(\mathbf{x}\right) - \mathbf{h}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})\right\}\right] = \mathbf{0}$$

- Mis-specify $E\left(\mathbf{g} \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}\right)$ does not cause inconsistency.

$$E\left[\left\{\mathbf{g}(Y, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}) - \mathbf{h}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})\right\}\left\{\boldsymbol{\alpha}\left(\mathbf{x}\right) - E\left(\boldsymbol{\alpha} \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}\right)\right\}\right] = \mathbf{0}$$

  This is extensively exploited in inverse regression type estimators.

- Cannot mis-specify both, but can estimate one or both nonparametrically.

Linear Algebra Tools: Two Lemmas

- Lemma 1: Assume $\mathbf{\Lambda}$ is a $p \times p$ symmetric matrix of rank d. If and only if β satisfies

$$\mathbf{\Lambda} - \mathbf{P}\mathbf{\Lambda}\mathbf{P} = \mathbf{0},$$

  then the span of the columns in β is the eigen-space of $\mathbf{\Lambda}$ corresponding to the d nonzero eigenvalues.
- Lemma 2: Assume $\mathbf{\Lambda}$ is a $p \times p$ symmetric non-negative definite matrix of rank d. If and only if β satisfies

$$\mathbf{Q}\mathbf{\Lambda}\mathbf{Q} = \mathbf{0},$$

  then the span of the columns in β is the eigen-space of $\mathbf{\Lambda}$ corresponding to the d nonzero eigenvalues.

SIR and Semi-SIR

- SIR
  - SIR estimate of $\mathcal{S}_{Y|\mathbf{x}}$: non-zero eigen-space of $\mathbf{\Lambda}_{\text{sir}} = \text{cov}\{E(\mathbf{x} \mid Y)\}$
  - SIR requirement: linearity condition
- Rederive SIR from semiparametric method
  - Set $\mathbf{g}(Y, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}) = E(\mathbf{x} \mid Y)$, $\boldsymbol{\alpha}(\mathbf{x}) = \mathbf{x}^{\mathrm{T}}$.
  - $E\{\boldsymbol{\alpha}(\mathbf{x}) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}\} = \mathbf{x}^{\mathrm{T}}\mathbf{P}$. Mis-specify $E\left(\mathbf{g} \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}\right) = 0$.
  - Result in $\mathbf{\Lambda}_{\text{sir}}\mathbf{Q} = \mathbf{0}$.
  - Lemma 2 ensures it is exactly SIR.
- Improve SIR by relaxing linearity condition, Semi-SIR:

$$E(\mathbf{x} \mid Y)\left\{\mathbf{x} - \widehat{E}(\mathbf{x} \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})\right\}^{\mathrm{T}}$$

SAVE and Semi-SAVE

- SAVE
  - SAVE estimate of $\mathcal{S}_{Y|\mathbf{x}}$: non-zero eigen-space of $\mathbf{\Lambda}_{\text{save}} = E\left[\{\mathbf{I}_p - \text{cov}(\mathbf{x} \mid Y)\}^2\right]$
  - SAVE restriction: linearity and constant variance condition
- Rederive SAVE from semiparametric method
  - $\mathbf{g}_1(Y, \beta^{\mathrm{T}}\mathbf{x}) = \mathbf{I}_p - \text{cov}(\mathbf{x} \mid Y)$, $\mathbf{g}_2(Y, \beta^{\mathrm{T}}\mathbf{x}) = \mathbf{g}_1 E(\mathbf{x} \mid Y)$, $\alpha_1(\mathbf{x}) = -\mathbf{x}\{\mathbf{x} - E(\mathbf{x} \mid \beta^{\mathrm{T}}\mathbf{x})\}^{\mathrm{T}}$, $\alpha_2(\mathbf{x}) = \mathbf{x}^{\mathrm{T}}$.
  - Calculate $E(\alpha_1 \mid \beta^{\mathrm{T}}\mathbf{x}), E(\alpha_2 \mid \beta^{\mathrm{T}}\mathbf{x})$. Mis-specify $E\left(\mathbf{g}_1 \mid \beta^{\mathrm{T}}\mathbf{x}\right) = E\left(\mathbf{g}_2 \mid \beta^{\mathrm{T}}\mathbf{x}\right) = \mathbf{0}$.
  - Result in $\mathbf{\Lambda}_{\text{save}}\mathbf{Q} = \mathbf{0}$.
  - Lemma 2 ensures it is exactly SAVE.
- Improve SAVE by relaxing both conditions, Semi-SAVE:

$$\{\mathbf{I}_p - \text{cov}(\mathbf{x} \mid Y)\}\left[\{\mathbf{x} - E(\mathbf{x} \mid Y)\}\left\{\mathbf{x} - \widehat{E}(\mathbf{x} \mid \beta^{\mathrm{T}}\mathbf{x})\right\}^{\mathrm{T}} - \widehat{\text{cov}}(\mathbf{x} \mid \beta^{\mathrm{T}}\mathbf{x})\right]$$

DR and Semi-DR

- DR
  - Non-zero eigen-space of $\Lambda_{\mathrm{dr}} = E[\{2\mathbf{I}_p - \mathbf{A}(Y, \widetilde{Y})\}^2]$, $\mathbf{A}(Y, \widetilde{Y}) = E\{(\mathbf{x} - \widetilde{\mathbf{x}})(\mathbf{x} - \widetilde{\mathbf{x}})^{\mathrm{T}} \mid Y, \widetilde{Y}\}$, $(\widetilde{\mathbf{x}}, \widetilde{Y}) \perp\!\!\!\perp (\mathbf{x}, Y)$.
  - DR requirement: linearity and constant variance condition
- Rederive DR from semiparametric method
  - $\mathbf{g}_1(Y, \beta^{\mathrm{T}}\mathbf{x}) = \mathbf{I}_p - E(\mathbf{x}\mathbf{x}^{\mathrm{T}} \mid Y)$,
    $\mathbf{g}_2(Y, \beta^{\mathrm{T}}\mathbf{x}) = E\left\{E(\mathbf{x} \mid Y)E(\mathbf{x}^{\mathrm{T}} \mid Y)\right\} E(\mathbf{x} \mid Y)$
    $\mathbf{g}_3(Y, \beta^{\mathrm{T}}\mathbf{x}) = E\left\{E(\mathbf{x}^{\mathrm{T}} \mid Y)E(\mathbf{x} \mid Y)\right\} E(\mathbf{x} \mid Y)$
    $\alpha_1(\mathbf{x}) = -\mathbf{x}\{\mathbf{x} - E(\mathbf{x} \mid \beta^{\mathrm{T}}\mathbf{x})\}^{\mathrm{T}}$, $\alpha_2(\mathbf{x}) = \alpha_3(\mathbf{x}) = \mathbf{x}^{\mathrm{T}}$.
  - Calculate $E(\alpha_i \mid \beta^{\mathrm{T}}\mathbf{x})$. Mis-specify $E(\mathbf{g}_i \mid \beta^{\mathrm{T}}\mathbf{x}) = \mathbf{0}$.
  - Result in $\Lambda_{\mathrm{dr}}\mathbf{Q} = \mathbf{0}$.
  - Lemma 2 ensures it is exactly DR.
- Improve DR by relaxing both conditions, Semi-DR

$$\left\{\mathbf{I}_p - E(\mathbf{x}\mathbf{x}^{\mathrm{T}} \mid Y)\right\}\left\{-\mathbf{x}\mathbf{x}^{\mathrm{T}} + \mathbf{x}\widehat{E}(\mathbf{x}^{\mathrm{T}} \mid \beta^{\mathrm{T}}\mathbf{x}) + \widehat{\mathrm{cov}}(\mathbf{x} \mid \beta^{\mathrm{T}}\mathbf{x})\right\}$$

$$+E\left\{E(\mathbf{x} \mid Y)E(\mathbf{x}^{\mathrm{T}} \mid Y)\right\} E(\mathbf{x} \mid Y)\left\{\mathbf{x}^{\mathrm{T}} - \widehat{E}(\mathbf{x}^{\mathrm{T}} \mid \beta^{\mathrm{T}}\mathbf{x})\right\}$$

$$+E\left\{E(\mathbf{x}^{\mathrm{T}} \mid Y)E(\mathbf{x} \mid Y)\right\} E(\mathbf{x} \mid Y)\left\{\mathbf{x}^{\mathrm{T}} - \widehat{E}(\mathbf{x}^{\mathrm{T}} \mid \beta^{\mathrm{T}}\mathbf{x})\right\}$$

Link to Non-elliptical SIR

- nSIR
  - nSIR estimate of $\mathcal{S}_{Y|\mathbf{x}}$: minimizing

$$E\left(\left\|E(\mathbf{x} \mid Y) - E\left\{E(\mathbf{x} \mid \beta^{\mathrm{T}}\mathbf{x}) \mid Y\right\}\right\|^2\right)$$

  - nSIR restriction: polynomial condition and constant variance condition
- Rederive nSIR from semiparametric method:
  - $\mathbf{a}(\mathbf{x}) = \mathbf{x} - E(\mathbf{x} \mid \beta^{\mathrm{T}}\mathbf{x})$
    $\mathbf{g}(Y, \beta^{\mathrm{T}}\mathbf{x}) = E\left[\partial E(\mathbf{x} \mid \beta^{\mathrm{T}}\mathbf{x})/\partial\left\{\mathrm{vec}(\beta)\right\}^{\mathrm{T}} \mid Y\right]$
  - $E(\mathbf{a} \mid \beta^{\mathrm{T}}\mathbf{x}) = \mathbf{0}$ guaranteed. Mis-specify $E(\mathbf{g} \mid \beta^{\mathrm{T}}\mathbf{x}) = \mathbf{0}$.
  - Result in nSIR.
- Improve nSIR by relaxing both conditions

Link to Non-elliptical SAVE

- nSAVE
  - nSAVE estimate of $\mathcal{S}_{Y|\mathbf{x}}$: minimizing

$$E\left(\left\|\mathbf{I}_p - \mathrm{cov}(\mathbf{x} \mid Y) - \mathrm{cov}\left\{E(\mathbf{x} \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})\right\} + \mathrm{cov}\left\{E(\mathbf{x} \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}) \mid Y\right\}\right\|^2\right)$$

  - Requirement: polynomial condition and constant variance condition
- Rederive nSAVE from semiparametric method
  - $\mathbf{g}(Y, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}) = \mathrm{vec}\left\{\mathbf{I}_p - \mathrm{cov}(\mathbf{x} \mid Y) - \mathrm{cov}\left(\mathbf{u}\right) + \mathrm{cov}\left(\mathbf{u} \mid Y\right)\right\}^{\mathrm{T}}$
    $\boldsymbol{\alpha}(\mathbf{x}) = \frac{\partial \mathrm{vec}(\mathbf{u}\mathbf{u}^{\mathrm{T}})}{\partial \mathrm{vec}(\boldsymbol{\beta})^{\mathrm{T}}} - \mathbf{u} \otimes \frac{\partial \mathbf{u}}{\partial \mathrm{vec}(\boldsymbol{\beta})^{\mathrm{T}}} - \frac{\partial \mathbf{u}}{\partial \mathrm{vec}(\boldsymbol{\beta})^{\mathrm{T}}} \otimes \mathbf{u}$
  - $E(\mathbf{g} \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}) = \mathbf{0}$ guaranteed. Mis-specify $E(\boldsymbol{\alpha} \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}) = \mathbf{0}$
  - Result in nSAVE.
- Improve nSAVE by relaxing both conditions

Semiparametric Estimation of $\mathcal{S}_{E(Y|\mathbf{x})}$

- Model: $E(Y \mid \mathbf{x}) = E(Y \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})$
- Equivalent form

$$Y = m(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}) + \epsilon, E(\epsilon \mid \mathbf{x}) = 0$$

- pdf: $\eta_1(\mathbf{x})\eta_2\{Y - m(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}), \mathbf{x}\}$, where $\int \epsilon\eta_2(\epsilon, \mathbf{x})d\mu(\epsilon) = 0$
- $\boldsymbol{\Lambda} = \boldsymbol{\Lambda}_1 \oplus \boldsymbol{\Lambda}_2$, where

$$\begin{aligned}
\boldsymbol{\Lambda}_1 &= \{\mathbf{f}(\mathbf{x}) : \forall \mathbf{f} \text{ such that } E(\mathbf{f}) = \mathbf{0}\} \\
\boldsymbol{\Lambda}_2 &= \{\mathbf{f}(\epsilon, \mathbf{x}) : \forall \mathbf{f} \text{ such that } E(\epsilon\mathbf{f} \mid \mathbf{x}) = E(\mathbf{f} \mid \mathbf{x}) = \mathbf{0}\}.
\end{aligned}$$

- $\boldsymbol{\Lambda}^{\perp} = \left[\{Y - E(Y \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})\}\left\{\boldsymbol{\alpha}(\mathbf{x}) - E\left(\boldsymbol{\alpha} \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}\right)\right\} : \forall \boldsymbol{\alpha}\right]$
- Fewer choices.
- Can still mis-specify either $E(Y \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})$ or $E(\boldsymbol{\alpha} \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})$.

OLS and Semi-OLS

- OLS
  - OLS estimate of $\mathcal{S}_{E(Y|\mathbf{x})}$: one subspace is $\mathrm{cov}(\mathbf{x}, Y)$
  - OLS requirement: linearity condition
- Rederive OLS from semiparametric method
  - Set $\alpha(\mathbf{x}) = \mathbf{x}$.
  - $E\{\alpha(\mathbf{x}) \mid \beta^{\mathrm{T}}\mathbf{x}\} = \mathbf{P}\mathbf{x}$. Mis-specify $E\left(Y \mid \beta^{\mathrm{T}}\mathbf{x}\right) = 0$
  - Obtain
    $\mathbf{0} = E(\mathbf{x}Y) - \mathbf{P}E(\mathbf{x}Y) = E(\mathbf{x}Y) - \beta(\beta^{\mathrm{T}}\beta)^{-1}\beta^{\mathrm{T}}E(\mathbf{x}Y)$
    This is exactly OLS.
- Improve OLS by relaxing linearity condition, Semi-OLS

$$Y\left\{\mathbf{x} - \widehat{E}(\mathbf{x} \mid \beta^{\mathrm{T}}\mathbf{x})\right\}$$

PHD and Semi-PHD

- PHD
  - PHD estimate of $\mathcal{S}_{E(Y|\mathbf{x})}$: non-zero eigen-space of $\mathbf{\Lambda}_{\text{phd}} = E\left[\{Y - E(Y)\}\,\mathbf{x}\mathbf{x}^{\mathrm{T}}\right]$
  - PHD requirement: linearity and constant variance condition
- Rederive PHD from semiparametric method
  - Set $\alpha(\mathbf{x}) = \mathbf{x}\mathbf{x}^{\mathrm{T}}$.
  - $\alpha(\mathbf{x}) - E\left\{\alpha(\mathbf{x}) \mid \beta^{\mathrm{T}}\mathbf{x}\right\} = \mathbf{x}\mathbf{x}^{\mathrm{T}} - \mathbf{Q} - \mathbf{P}\mathbf{x}\mathbf{x}^{\mathrm{T}}\mathbf{P}$
    Mis-specify $E\left(Y \mid \beta^{\mathrm{T}}\mathbf{x}\right) = E(Y)$.
  - Obtain
    $E\left[\{Y - E(Y)\}\left(\mathbf{x}\mathbf{x}^{\mathrm{T}} - \mathbf{P}\mathbf{x}\mathbf{x}^{\mathrm{T}}\mathbf{P}\right)\right] = \mathbf{\Lambda}_{\text{phd}} - \mathbf{P}\mathbf{\Lambda}_{\text{phd}}\mathbf{P} = \mathbf{0}$
  - Lemma 1 ensures it is exactly PHD
- Improve PHD by relaxing both conditions, Semi-PHD

$$\{Y - E(Y)\}\left\{\mathbf{x}\mathbf{x}^{\mathrm{T}} - \widehat{E}(\mathbf{x}\mathbf{x}^{\mathrm{T}} \mid \beta^{\mathrm{T}}\mathbf{x})\right\}$$

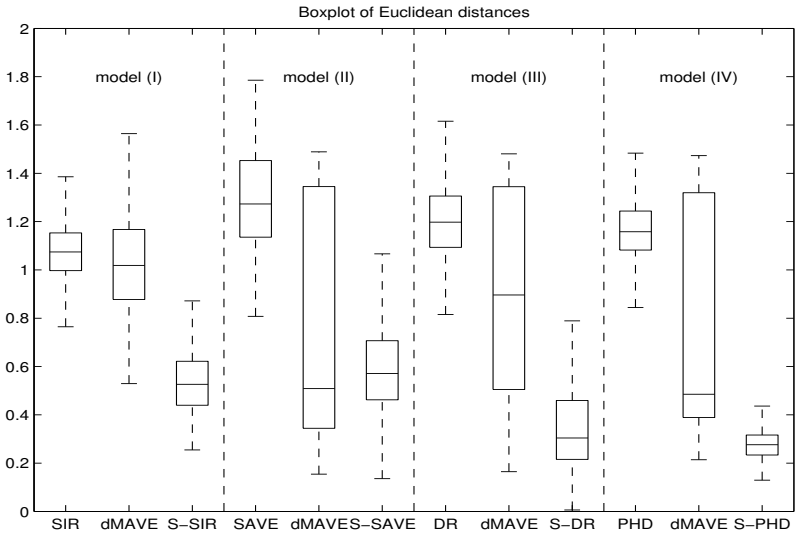
Simulation

- $p = 12, d = 2, n = 200, 500$ simulations
- Four models:

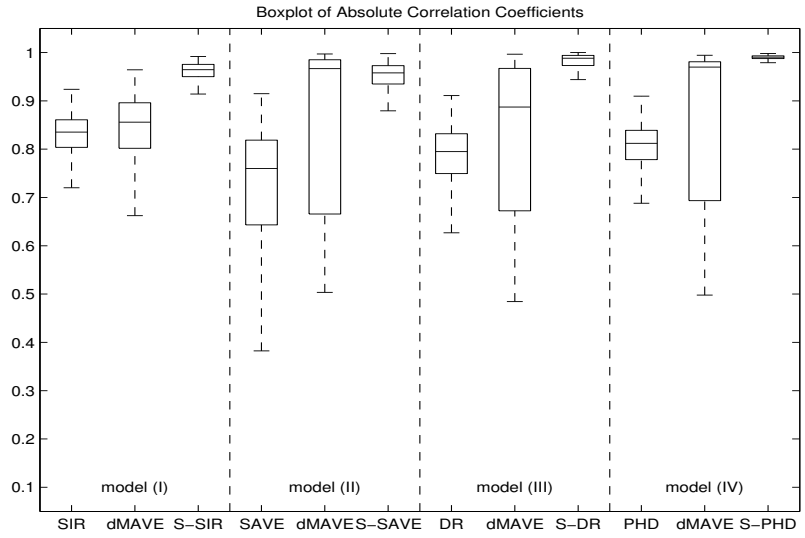
$$\begin{aligned}
\text{model 1}: &\quad Y = \left(\mathbf{x}^{\mathrm{T}}\beta_1\right) \Big/ \left\{0.5 + \left(\mathbf{x}^{\mathrm{T}}\beta_2 + 1.5\right)^2\right\} + 0.5\varepsilon; \\
\text{model 2}: &\quad Y = \left(\mathbf{x}^{\mathrm{T}}\beta_1\right)^2 + 2 \mid \mathbf{x}^{\mathrm{T}}\beta_2 \mid + 0.1 \mid \mathbf{x}^{\mathrm{T}}\beta_2 \mid \varepsilon; \\
\text{model 3}: &\quad Y = \exp\left(\mathbf{x}^{\mathrm{T}}\beta_1\right) + 2\left(\mathbf{x}^{\mathrm{T}}\beta_2 + 1\right)^2 + \mid \mathbf{x}^{\mathrm{T}}\beta_1 \mid \varepsilon; \\
\text{model 4}: &\quad Y = \left(\mathbf{x}^{\mathrm{T}}\beta_1\right)^2 + \left(\mathbf{x}^{\mathrm{T}}\beta_2\right)^2 + 0.5\varepsilon,
\end{aligned}$$

- Two cases:
  case 1: $\mathbf{x}$ does not satisfy linearity or constant variance condition
  case 2: $\mathbf{x}$ satisfies both linearity and constant variance condition

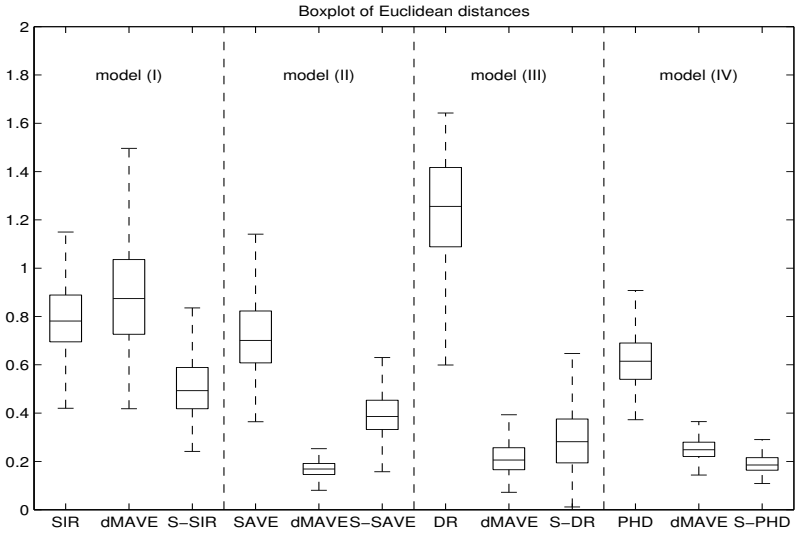
Simulation Results and a Paradox



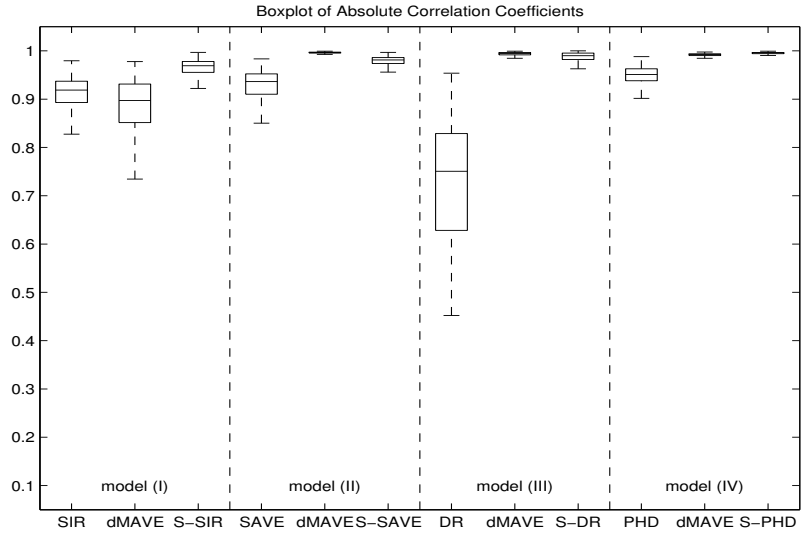
case 1: Euclidean distances



case 1: Canonical correlations



case 2: Euclidean distances



case 2: Canonical correlations

Inference and Parameterization

- Inference on space estimation done via inference on parameter estimation
- Parameterization (up to a zero measure set)
  - $\boldsymbol{\beta}$ is an orthogonal matrix: not unique
  - Only for $d = 1$: $\beta_k = \sin(\theta_1) \ldots \sin(\theta_{k-1}) \cos(\theta_k)$ for $k < p$, $\beta_p = \sin(\theta_1) \ldots \sin(\theta_{p-2}) \sin(\theta_{p-1})$: computation cumbersome
  - Let $\boldsymbol{\beta}_l \in \mathcal{R}^{(p-d)\times d}$ completely free,

$$\boldsymbol{\beta} = \begin{pmatrix} \mathbf{I}_d \\ \boldsymbol{\beta}_l \end{pmatrix}$$

  - If $\boldsymbol{\beta}$ spans the dimension reduction space, so does $\boldsymbol{\beta}\mathbf{A}$, $\forall \mathbf{A} \in \mathcal{R}^{d\times d}$ and invertable.
  - A parameterization of a point on the Grassmann manifold (collection of d-dimensional subspaces of p-dimensional real space)
- $\mathrm{vecl}(\mathbf{A})$: vectorize lower $(p-d) \times d$ submatrix of $\mathbf{A}$

Inference in Semiparametric Estimation

- Reducing the number of estimating equations to $(p-d)d$ via GMM
- Standard asymptotic analysis to obtain large sample properties on $\widehat{\beta}_I$
- Typically, consistent estimation in weight matrix has no first order effect.
- Typically, if properly done, starting from a function in $\Lambda^{\perp}$, additional nonparametric estimation has no first order effect.

Is the Paradox Real?

- Recall $\{\mathbf{g}(Y, \beta^{\mathrm{T}}\mathbf{x}) - E(\mathbf{g} \mid \beta^{\mathrm{T}}\mathbf{x})\}\{\mathbf{a}(\mathbf{x}) - E(\mathbf{a} \mid \beta^{\mathrm{T}}\mathbf{x})\}$
- Inverse regression methods solve

$$\sum_{i=1}^{n}\{\mathbf{g}(Y_i, \beta^{\mathrm{T}}\mathbf{x}_i) - h(\beta^{\mathrm{T}}\mathbf{x}_i)\}\{\mathbf{a}(\mathbf{x}_i) - E(\mathbf{a} \mid \beta^{\mathrm{T}}\mathbf{x}_i)\} = \mathbf{0}$$

- Semiparametric estimators solve

$$\sum_{i=1}^{n}\{\mathbf{g}(Y_i, \beta^{\mathrm{T}}\mathbf{x}_i) - h(\beta^{\mathrm{T}}\mathbf{x}_i)\}\{\mathbf{a}(\mathbf{x}_i) - \widehat{E}(\mathbf{a} \mid \beta^{\mathrm{T}}\mathbf{x}_i)\} = \mathbf{0}$$

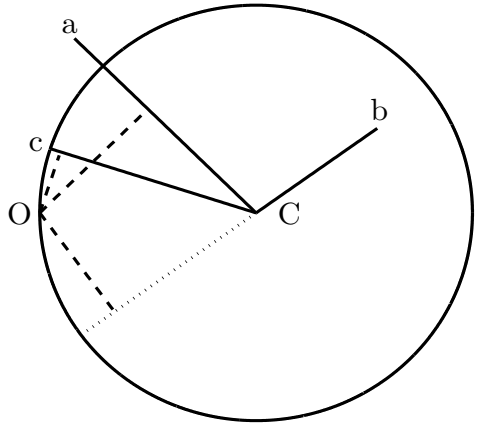
- It is real! Formal asymptotic analysis shows semiparametric estimator more efficient than inverse regression estimator.

Intuitive Understanding

- General parametric models: $\mathbf{a}(\mathbf{x}) = \mathbf{m}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}, \boldsymbol{\alpha}) + \epsilon$
  Solve $\sum_{i=1}^{n} \mathbf{A}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)\{\mathbf{a}(\mathbf{x}_i) - \mathbf{m}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i, \boldsymbol{\alpha})\} = \mathbf{0}$ to obtain $\widehat{\boldsymbol{\alpha}}(\boldsymbol{\beta})$.
- Three estimators

$$
\begin{aligned}
(L) \quad & \textstyle\sum_{i=1}^{n} \mathbf{g}(Y_i, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)\{\mathbf{a}(\mathbf{x}_i) - \mathbf{m}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)\} = \mathbf{0}, \\
(G) \quad & \textstyle\sum_{i=1}^{n} \mathbf{g}(Y_i, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)\{\mathbf{a}(\mathbf{x}_i) - \mathbf{m}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i, \widehat{\boldsymbol{\alpha}}(\boldsymbol{\beta})\} = \mathbf{0}, \\
(N) \quad & \textstyle\sum_{i=1}^{n} \mathbf{g}(Y_i, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)\{\mathbf{a}(\mathbf{x}_i) - \widehat{\mathbf{m}}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)\} = \mathbf{0}.
\end{aligned}
$$

- $\mathrm{var}(L) = \mathrm{var}(N) + \mathrm{var}(C), \mathrm{var}(G) = \mathrm{var}(N) + \mathrm{var}(a/b/c)$. It is not true that $L > G > N$.

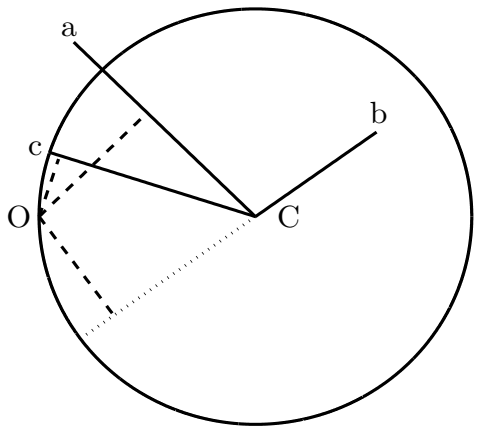


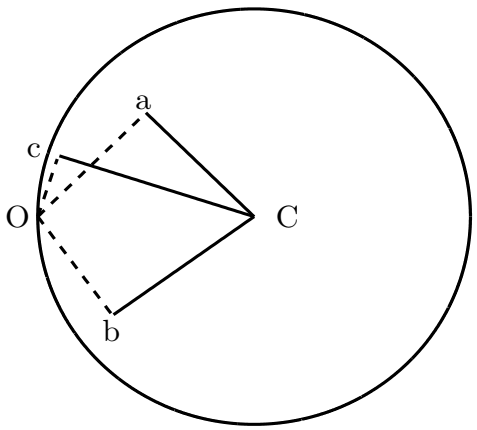
Under OWLS

- OWLS:
  $\sum_{i=1}^{n} \mathbf{m}'(\beta^{\mathrm{T}}\mathbf{x}_i, \alpha)\mathbf{V}^{-1}(\beta^{\mathrm{T}}\mathbf{x}_i)\{\mathbf{a}(\mathbf{x}_i) - \mathbf{m}(\beta^{\mathrm{T}}\mathbf{x}_i, \alpha)\} = \mathbf{0}$
- Three estimators

$$
\begin{aligned}
(L) \quad & \textstyle\sum_{i=1}^{n} \mathbf{g}(Y_i, \beta^{\mathrm{T}}\mathbf{x}_i)\{\mathbf{a}(\mathbf{x}_i) - \mathbf{m}(\beta^{\mathrm{T}}\mathbf{x}_i)\} = \mathbf{0}, \\
(OG) \quad & \textstyle\sum_{i=1}^{n} \mathbf{g}(Y_i, \beta^{\mathrm{T}}\mathbf{x}_i)\{\mathbf{a}(\mathbf{x}_i) - \mathbf{m}(\beta^{\mathrm{T}}\mathbf{x}_i, \widehat{\alpha}(\beta)\} = \mathbf{0}, \\
(N) \quad & \textstyle\sum_{i=1}^{n} \mathbf{g}(Y_i, \beta^{\mathrm{T}}\mathbf{x}_i)\{\mathbf{a}(\mathbf{x}_i) - \widehat{\mathbf{m}}(\beta^{\mathrm{T}}\mathbf{x}_i)\} = \mathbf{0}.
\end{aligned}
$$

- $\mathrm{var}(L) = \mathrm{var}(N) + \mathrm{var}(C), \mathrm{var}(OG) = \mathrm{var}(N) + \mathrm{var}(a/b/c)$.
  Always $L > OG > N$.





Nested Model and Best Estimator

- For nested models, $a/b/c$ becomes distance to a line, a plain, a hyperplain...
- As dimension of the hyperplain increases, the distance decreases.
- Nonparametric kernel estimation is the extreme case of OWLS.
- This is the geometric intuition of why ignoring linearity/constant variance leads to better estimator.

A Fair Evaluation of Linearity Condition

- Inverse regression methods rely on Linearity/Constant Variance condition to be consistent, the price is inflation of variability
- Linearity/Constant Variance condition does not bring loss if not overused
- Linearity/Constant Variance condition brings no gain if not taken into account at beginning
- Optimal use of these conditions improves efficiency compared to without these conditions. Efficient estimator forms very complex.

Efficient Estimation of $\mathcal{S}_{Y|\mathbf{x}}$

- Score function: $\mathbf{S}_{\beta} = \text{vecl}\{\mathbf{x}\partial\log\eta_2(Y, \beta^{\mathrm{T}}\mathbf{x})/\partial(\mathbf{x}^{\mathrm{T}}\beta)\}$
- Efficient score

$$\mathbf{S}_{\text{eff}}(Y, \mathbf{x}, \beta^{\mathrm{T}}\mathbf{x}, \eta_2, E) = \text{vecl}\left[\{\mathbf{x} - E(\mathbf{x} \mid \beta^{\mathrm{T}}\mathbf{x})\}\frac{\partial\log\eta_2(Y, \beta^{\mathrm{T}}\mathbf{x})}{\partial(\mathbf{x}^{\mathrm{T}}\beta)}\right]$$

- Estimate $\eta_2, \partial\eta_2/\partial(\mathbf{x}^{\mathrm{T}}\beta), E(\cdot \mid \beta^{\mathrm{T}}\mathbf{x})$: all d-dimensional problems
- Efficient estimation

$$\sum_{i=1}^{n} \mathbf{S}_{\text{eff}}(Y_i, \mathbf{x}_i, \beta^{\mathrm{T}}\mathbf{x}_i, \widehat{\eta}_2, \widehat{\eta}_2', w, \widehat{E}) = \mathbf{0}$$

- $\sqrt{n}\text{vecl}(\widehat{\beta} - \beta) \to N[\mathbf{0}, \{\mathbf{S}_{\text{eff}}^{\otimes 2}(Y, \mathbf{x}, \beta^{\mathrm{T}}\mathbf{x}, \eta_2, E)\}^{-1}]$.

Efficient Estimation of $\mathcal{S}_{E(Y|\mathbf{x})}$

- Score function

$$\mathbf{S}_\beta = -\text{vecl}\left\{\frac{\mathbf{x}\eta'_{2\epsilon}(\epsilon,\mathbf{x})\partial m(\beta^{\mathrm{T}}\mathbf{x})}{\eta_2(\epsilon,\mathbf{x})\partial(\mathbf{x}^{\mathrm{T}}\beta)}\right\}$$

- Efficient estimation of $\mathcal{S}_{E(Y|\mathbf{x})}$. Let $w(\mathbf{x}) = 1/E(\epsilon^2 \mid \mathbf{x})$.

$$\begin{aligned} & \mathbf{S}_{\text{eff}}(Y,\mathbf{x},\beta^{\mathrm{T}}\mathbf{x},m,E,w) \\ = \quad & \text{vecl}\left(\epsilon w(\mathbf{x})\left[\mathbf{x} - \frac{E\{\mathbf{x}w(\mathbf{x}) \mid \beta^{\mathrm{T}}\mathbf{x}\}}{E\{w(\mathbf{x}) \mid \beta^{\mathrm{T}}\mathbf{x}\}}\right]\frac{\partial m(\beta^{\mathrm{T}}\mathbf{x})}{\partial(\mathbf{x}^{\mathrm{T}}\beta)}\right) \end{aligned}$$

- Estimate $m, m', E(\cdot \mid \beta^{\mathrm{T}}\mathbf{x})$: all d-dimensional problems
- Estimate w: p-dimensional problem, consistency suffices
- Efficient estimation

$$\textstyle\sum_{i=1}^n \mathbf{S}_{\text{eff}}(Y_i,\mathbf{x}_i,\beta^{\mathrm{T}}\mathbf{x}_i,\widehat{m},\widehat{m}',\widehat{E}(\cdot \mid \beta^{\mathrm{T}}\mathbf{x}),\widehat{w}) = \mathbf{0}$$

- $\sqrt{n}\text{vecl}(\widehat{\beta} - \beta) \to N[\mathbf{0}, \{\mathbf{S}_{\text{eff}}^{\otimes 2}(Y,\mathbf{x},\beta^{\mathrm{T}}\mathbf{x},m,E,w)\}^{-1}]$

Simulation $\mathcal{S}_{Y|\mathbf{x}}$

		β_1 1.3	β_2 -1.3	β_3 1	β_4 -.5	β_5 .5	β_6 -.5
Ora	ave	1.2978	-1.3036	1.0049	-0.4985	0.5033	-0.4943
	std	0.1221	0.1477	0.1505	0.1169	0.0966	0.1049
	$\widehat{\text{std}}$	0.1264	0.1510	0.1527	0.1212	0.0983	0.1052
	95%	0.9510	0.9540	0.9440	0.9540	0.9520	0.9450
Eff	ave	1.2980	-1.3046	1.0064	-0.4990	0.5040	-0.4936
	std	0.1280	0.1546	0.1567	0.1221	0.1000	0.1075
	$\widehat{\text{std}}$	0.1317	0.1588	0.1602	0.1264	0.1011	0.1084
	95%	0.9480	0.9380	0.9380	0.9440	0.9480	0.9510
Loc	ave	1.3052	-1.2629	0.9687	-0.4988	0.5023	-0.4897
	std	0.1478	0.1736	0.1715	0.1393	0.1069	0.1153
dMA	ave	1.2599	-1.2933	1.0014	-0.4763	0.4984	-0.4935
	std	0.1932	0.1427	0.1550	0.1701	0.1368	0.1378
SIR	ave	1.3881	-1.1930	0.9261	-0.5968	0.4793	-0.4724
	std	0.1696	0.1522	0.1414	0.1489	0.0976	0.0995
DR	ave	0.9935	-0.2217	0.1930	-0.6863	0.1245	-0.1071
	std	0.6567	1.2305	1.0107	0.6411	0.3069	0.2999

Simulation $\mathcal{S}_{E(Y|\mathbf{x})}$

		β_1-1	β_2-2	β_3-0.5	β_4-1	β_5-2	β_6-2	β_7-1.5	β_8-1
o	$\widehat{\beta}$	-1.03	-2.08	-0.52	-1.04	-2.05	-2.05	-1.55	-1.02
r	σ	0.11	0.14	0.08	0.08	0.14	0.16	0.11	0.10
c	CI	95.6	93.0	93.9	94.2	94.9	94.2	93.6	94.4
l	$\widehat{\beta}$	-1.04	-2.09	-0.52	-1.05	-2.06	-2.05	-1.55	-1.03
o	σ	0.11	0.14	0.08	0.08	0.15	0.17	0.11	0.10
1	CI	94.6	92.4	94.1	92.9	93.2	93.2	92.8	92.7
l	$\widehat{\beta}$	-1.02	-2.05	-0.51	-1.03	-2.04	-2.04	-1.54	-1.02
o	σ	0.11	0.15	0.08	0.09	0.16	0.18	0.12	0.11
2	CI	94.9	94.0	94.8	94.1	94.1	94.6	93.6	92.8
l	$\widehat{\beta}$	-1.03	-2.07	-0.51	-1.04	-2.05	-2.04	-1.54	-1.02
o	σ	0.11	0.14	0.08	0.08	0.14	0.16	0.11	0.10
3	CI	95.1	94.0	94.7	93.5	95.2	94.2	94.5	94.3
M	$\widehat{\beta}$	-1.01	-1.88	-0.52	-1.06	-1.91	-1.86	-1.48	-1.01
A	σ	0.17	0.20	0.13	0.12	0.22	0.26	0.18	0.16
S	$\widehat{\beta}$	-1.00	-1.92	-0.51	-1.03	-2.03	-1.94	-1.54	-1.04
P	σ	0.17	0.19	0.12	0.15	0.36	0.34	0.27	0.30

Determine d in Inverse Regression Based Methods

- Inverse regression based method procedure
  - Estimate $\mathbf{\Lambda}$ from data, get $\widehat{\mathbf{\Lambda}}$.
  - Estimate the nonzero eigenspace of $\widehat{\mathbf{\Lambda}}$.
- d is the number of non-zero eigenvalues of the matrix $\mathbf{\Lambda}$.
- Deciding d is to determine the number of non-zero eigenvalues of $\mathbf{\Lambda}$ based on $\widehat{\mathbf{\Lambda}}$.

Sequential Test

- Sequential test procedure:
  1. H_{01} : $\mathbf{\Lambda}$ has $p-1$ non-zero eigenvalues, if H_{01} passes, $\widehat{d} = p-1$, otherwise
  2. H_{02} : $\mathbf{\Lambda}$ has $p-2$ non-zero eigenvalues, if H_{02} passes, $\widehat{d} = p-2$, otherwise
  3. ...
- Sequential test problems:
  - $\widehat{d}$ not consistent due to type-I errors allowed.
  - Cumulative type-I errors may be not ignorable.
  - When p large, computationally inefficient.
  - Heavily relies on normality of $\widehat{\mathbf{\Lambda}}$.
  - Needs $\mathrm{cov}(\widehat{\mathbf{\Lambda}})$.
  - Done case-by-case for each inverse regression based estimator.

BIC

- BIC procedure: The goodness-of-fit of a rank k matrix $\mathbf{\Lambda}_k$ to $\widehat{\mathbf{\Lambda}}$, $f(\widehat{\mathbf{\Lambda}}, \mathbf{\Lambda}_k)$ is penalized by the rank k, $c(n, k)$.
- General form $\min_k f(\widehat{\mathbf{\Lambda}}, \mathbf{\Lambda}_k) + c(n, k)$
- BIC properties and problems:
  - It works with $\widehat{\mathbf{\Lambda}} + \mathbf{I}_p$ and $\mathbf{\Lambda}_k + \mathbf{I}_p$, instead of $\widehat{\mathbf{\Lambda}}$ and $\mathbf{\Lambda}_k$ for technical reason.
  - Construction of goodness-of-fit f based on normality of $\widehat{\mathbf{\Lambda}}$.
  - Designed only for SIR.
  - It is consistent as long as $\widehat{\mathbf{\Lambda}}$ is consistent.
  - It is sensitive to the amount of the penalty $c(n, k)$.

Sparse Eigen-Decomposition (SED)

- Eigen decomposition problem is least square problem:

$$\min_{\lambda_i, \boldsymbol{\xi}_i, \boldsymbol{\eta}_i} \|\boldsymbol{\Lambda} - \sum_{i=1}^{p} \lambda_i \boldsymbol{\xi}_i \boldsymbol{\eta}_i^{\mathrm{T}}\|,$$
$$\text{where } \|\boldsymbol{\xi}_i\| = \|\boldsymbol{\eta}_i\| = 1, \lambda_1 \geq \cdots \geq \lambda_p.$$

- Equivalently minimize

$$\|\boldsymbol{\Lambda} - \sum_{i=1}^{p} \lambda_i \boldsymbol{\xi}_i \boldsymbol{\eta}_i^{\mathrm{T}}\|^2 = \|\boldsymbol{\Lambda}\|^2 - 2\sum_{i=1}^{p} \lambda_i \boldsymbol{\xi}_i^{\mathrm{T}} \boldsymbol{\Lambda} \boldsymbol{\eta}_i + \sum_{i=1}^{p} \lambda_i^2$$

  to obtain $\widehat{\boldsymbol{\xi}}_i(\lambda'\mathrm{s}), \widehat{\boldsymbol{\eta}}_i(\lambda'\mathrm{s})$.

- SED procedure (least square+lasso)

$$\min_{\lambda'\mathrm{s}} \|\boldsymbol{\Lambda} - \sum_{i=1}^{p} \lambda_i \widehat{\boldsymbol{\xi}}_i(\lambda'\mathrm{s}) \widehat{\boldsymbol{\eta}}_i(\lambda'\mathrm{s})^{\mathrm{T}}\| + c(n) \sum_{i=1}^{p} \widehat{w}_i |\lambda_i|$$

- SED property: $\widehat{d}$ and $\widehat{\lambda}'$s consistent
- LARS applicable, computationally simple

Bootstrap

- Iintuition
  - Under true d, estimated space $\mathbf{S}(d)$ approximates true space, has variability due to data randomness
  - Under $k > d$, estimated space $\mathbf{S}(k)$ contains zero-eigendirection. Zero-eigendirection is not unique in population, the choice of zero-eigendirection in finite sample inflates variability
  - Under $k < d$, estimated space $\mathbf{S}(k)$ is a subspace of $\mathbf{S}(d)$. The choice of the subspace in finite sample inflates variability.
- Variability of space estimation assessed via bootstrap
- Theoretical justification not available

Determine d in Nonparametric Methods

- Crossvalidation procedure
  - $\mathcal{S}_{E(Y|\mathbf{x})}$

$$\mathrm{CV}(k) = n^{-1} \sum_{i=1}^{n} \{Y_i - \widehat{m}_{-i}(\widehat{\boldsymbol{\beta}}_{-i}^{\mathrm{T}} \mathbf{x}_i)\}^2$$

  - $\mathcal{S}_{Y|\mathbf{x}}$

$$\mathrm{CV}(k) = n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} \{K_b(Y_i - Y_j) - \widehat{\eta}_{2,-i}(Y_j, \widehat{\boldsymbol{\beta}}_{-i}^{\mathrm{T}} \mathbf{x}_i)\}^2$$

- Requires a goodness-of-fit criterion, leads to estimating quantities that could be avoided in other methods
- Computationally demanding

Determine d in Semiparametric Methods

- Estimating equation based estimation.
- Cannot be transformed to a problem of testing zero eigenvalues without linearity condition.
- No likelihood or other criterion computable, so AIC, BIC, CV do not apply.
- Quadratic form cannot replace likelihood as a criterion.
- Not a problem of zero coefficient: penalty does not apply.

Bootstrap

- Intuition under true d
  - Data based $\widehat{\mathbf{S}}(d)$ approximates true space $\mathbf{S}(d)$
  - Bootstrap data based $\widehat{\mathbf{S}}_B(d)$ approximates true space $\mathbf{S}(d)$
  - $\widehat{\mathbf{S}}(d)$ and $\widehat{\mathbf{S}}_B(d)$ highly correlated
- Intuition under $k > d$
  - Data based $\widehat{\mathbf{S}}(k)$ contains arbitrary direction out of $\mathbf{S}(d)$
  - Bootstrap data based $\widehat{\mathbf{S}}_B(k)$ contains arbitrary direction out of $\mathbf{S}(d)$
  - Arbitrary direction determined by data, $\widehat{\mathbf{S}}(k)$ and $\widehat{\mathbf{S}}_B(k)$ not as highly correlated
- Intuition under $k < d$
  - Data based $\widehat{\mathbf{S}}(k)$ is arbitrary subspace of $\mathbf{S}(d)$
  - Bootstrap data based $\widehat{\mathbf{S}}_B(k)$ is arbitrary subspace of $\mathbf{S}(d)$
  - Arbitrary subspace determined by data, $\widehat{\mathbf{S}}(k)$ and $\widehat{\mathbf{S}}_B(k)$ not as highly correlated
- No theoretical proof for its validity

VIC

- Key observation in semiparametric estimating equations
  1. If $k < d$, adding bad direction destroys estimating equation consistency under $k+1$ model.
  2. If $k \geq d$, adding bad direction does not destroy estimating equation consistency under $k+1$ model.
- VIC stands for Validated Information Criterion.
- Main idea: Extend a current β estimator by adding a column, examine if consistency is destroyed or not. Stop as soon as consistency does not get destroyed.
- We might get too lucky. Extend $\widehat{\beta}_{(k)}$ to $\widehat{\gamma}^1_{(k)}, \widehat{\gamma}^2_{(k)}$.

$$\begin{aligned} & \mathrm{VIC}(k) \\ = \; & \frac{1}{2n}\left\{ \| \sum_{i=1}^{n} \mathbf{f}(\mathbf{O}_i, \widehat{\gamma}^1_{(k)}) \|^2 + \| \sum_{i=1}^{n} \mathbf{f}(\mathbf{O}_i, \widehat{\gamma}^2_{(k)}) \|^2 \right\} + pk\log(n) \end{aligned}$$

- $\widehat{d} = \mathrm{argmin}_k \; \mathrm{VIC}(k)$

Properties of VIC

- Only need to solve for β in smaller (k) model. No equation solving in larger $(k+1)$ model.
- Theorem on selection consistency:
  Let $\widehat{d} = \text{argmin}_k \text{VIC}(k)$. Under regularity conditions, when $n \to \infty$, $\text{pr}(\widehat{d} = d) \to 1$.
- Simulation: repeated 1000 times. Conducted for $n = 200, n = 400$, $p = 6, 10$ and $d = 1, 2$. Performed estimation after selecting structural dimension.

Simulation Results: $d = 1$

Selection frequency by VIC(k) in %, $p = 6$.

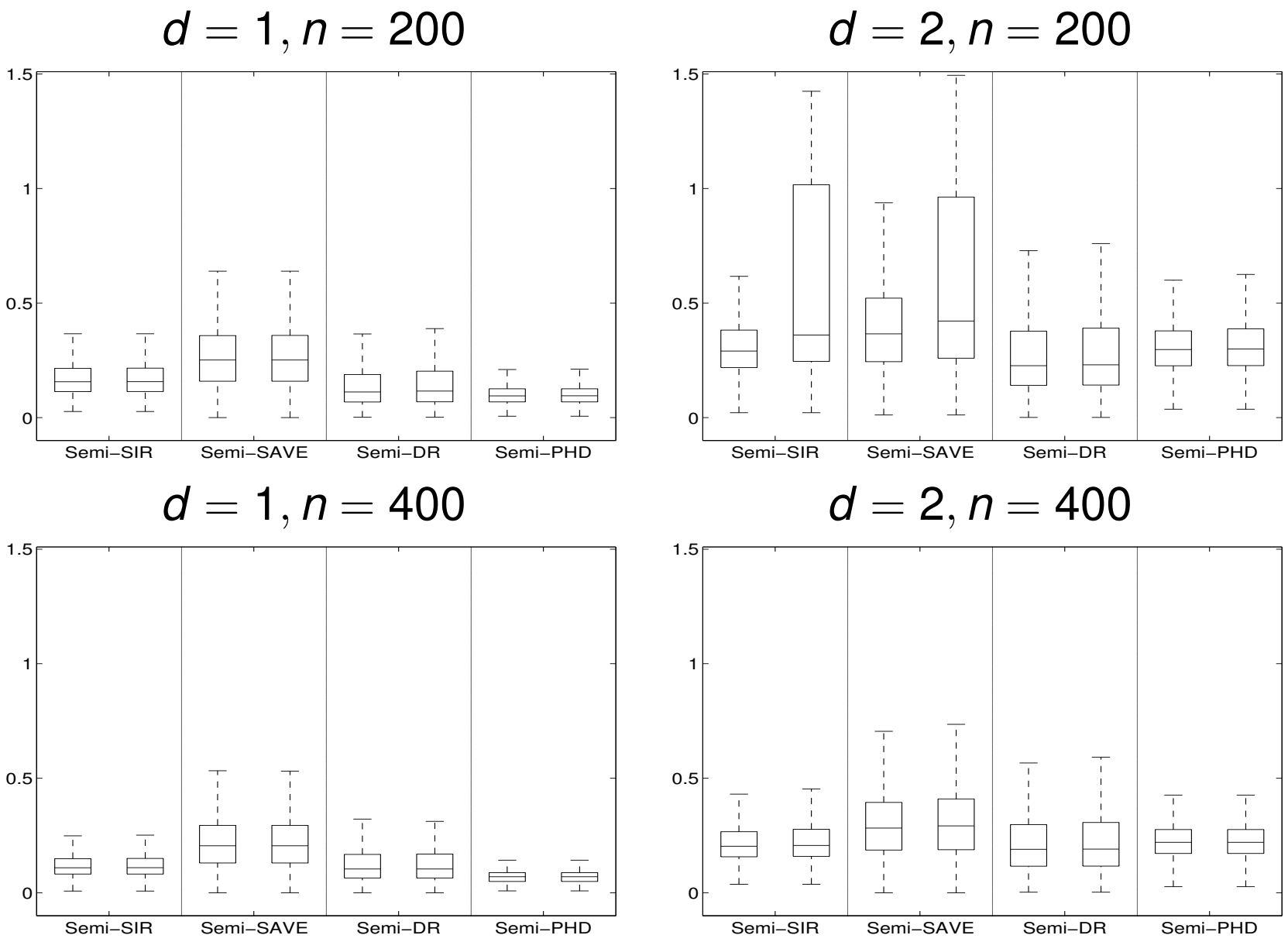
	$\widehat{d} = 1$	$\widehat{d} = 2$	$\widehat{d} = 3$	$\widehat{d} = 4$
		$n = 200$		
Semi-SIR	99.4	0.6	0.0	0.0
Semi-SAVE	94.1	5.2	0.6	0.1
Semi-DR	85.6	14.0	0.4	0.0
Semi-PHD	99.9	0.1	0.0	0.0
		$n = 400$		
Semi-SIR	99.7	0.3	0.0	0.0
Semi-SAVE	94.4	5.5	0.1	0.0
Semi-DR	86.8	12.8	0.4	0.0
Semi-PHD	99.9	0.1	0.0	0.0

Simulation Results: $d = 2$

Selection frequency by VIC(k) in %, $p = 6$.

	$\widehat{d} = 1$	$\widehat{d} = 2$	$\widehat{d} = 3$	$\widehat{d} = 4$
		$n = 200$		
Semi-SIR	31.0	68.3	0.6	0.1
Semi-SAVE	9.9	76.0	12.1	2.0
Semi-DR	1.4	96.6	1.7	0.3
Semi-PHD	2.6	97.2	0.2	0.0
		$n = 400$		
Semi-SIR	4.6	94.9	0.5	0.0
Semi-SAVE	1.7	91.7	5.7	0.9
Semi-DR	1.6	98.1	0.2	0.1
Semi-PHD	0.0	99.8	0.2	0.0

Estimation Results ($p = 6$, true and selected d)



Generality and Extension of VIC

- Supplement to AIC, BIC.
- Can be used in any situation when estimating equation exists, but criterion not available or hard to compute: No AIC, BIC, CV.
  - moment based methods
  - GEE
  - Martingale based methods
  - semiparametric methods
- Especially useful when model is nested, but smaller model is not result of having zero values in the larger model. No test, penalization.
- Especially useful when increasing model complexity implies great elevation of computation.

When p Is Very Large

- SIR valid for $p = o(n^{1/4})$
- Cumulative slicing valid for $p = o(n^{1/2})$
- Hypotheses
  - All current methods hold for $p = o(n^{1/2})$
  - All current methods fail for $p \neq o(n^{1/2})$
- Challenges and suggestions for $p >> n^{1/2}$
  - $\widehat{\Sigma}_x$ singular, normalization no longer apply
  - $\widehat{\Sigma}_x^{-1}$ not available, avoid it by partial least square idea
- Untra large p requires additional assumptions
  - Screening
  - Assuming sparsity, performing penalizing
  - Introducing latent structure

Extensions

- Extension to nonlinear case: $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x} \Longrightarrow \mathbf{u}(\mathbf{x}, \boldsymbol{\beta})$
- Extension to nonparametric case: $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x} \Longrightarrow \mathbf{u}(\mathbf{x})$
- Extension to functional dimension reduction
- Extension to two-way and multi-way dimension reduction (dimension folding)
- Extension to multivariate response
- Extension to incorporating additional constraints (envelope models)